

Anonymous Networking amidst Eavesdroppers

Parvathinathan Venkitasubramaniam, Ting He and Lang Tong

School of Electrical and Computer Engineering

Cornell University, Ithaca, NY 14853

Email : {pv45, th255, lt35}@cornell.edu

Abstract—The problem of security against timing based traffic analysis in wireless networks is considered in this work. An analytical measure of anonymity in eavesdropped networks is proposed using the information theoretic concept of equivocation. For a physical layer with orthogonal transmitter directed signaling, scheduling and relaying techniques are designed to maximize achievable network performance for any given level of anonymity. The network performance is measured by the achievable relay rates from the sources to destinations under latency and medium access constraints. In particular, analytical results are presented for two scenarios:

For a two-hop network with maximum anonymity, achievable rate regions for a general $m \times 1$ relay are characterized when nodes generate independent Poisson transmission schedules. The rate regions are presented for both strict and average delay constraints on traffic flow through the relay.

For a multihop network with an arbitrary anonymity requirement, the problem of maximizing the sum-rate of flows (network throughput) is considered. A selective independent scheduling strategy is designed for this purpose, and using the analytical results for the two-hop network, the achievable throughput is characterized as a function of the anonymity level. The throughput-anonymity relation for the proposed strategy is shown to be equivalent to an information theoretic rate-distortion function.

Index Terms—Network Security, Traffic Analysis, Secrecy, Rate-Distortion.

I. INTRODUCTION

Traffic analysis attacks are carried out by eavesdroppers monitoring node transmissions to obtain networking information such as source-destination pairs and paths of data flow. Traffic analysis has played a prominent role in modern warfare [1] and its adverse effects on computer networks is well documented in literature [2], [3], [4], [5]. For example, the weaknesses of protocols for web browsing [4], [6] and SSH [7] have been exposed through traffic analysis.

The primary focus of this work is an analytical approach to security against traffic analysis in wireless networks and the design of provably secure countermeasures. Owing to the unprotected medium of communication, eavesdropping node transmissions in wireless networks is easy and undetectable. Although cryptography can be used to prevent analysis based on contents or packet lengths (see Section I-B), the knowledge of transmission epochs alone can reveal critical information such as paths of information flow. We address the problem

of designing *anonymous* transmission schedules and relaying strategies to counter the transmission epoch based inference of data flows by eavesdroppers.

The challenge in designing *anonymous* transmission strategies is to adhere to the networking constraints while hiding information from eavesdroppers. Wireless networks are subject to constraints on medium access, latency and stability, which generally result in a high correlation across transmission schedules of nodes in a path. The need for anonymity however necessitates that paths are not revealed by correlation of transmission schedules. These contrasting paradigms result in a tradeoff between anonymity and network performance. For example, consider the simple two hop setup shown in Fig. 1, wherein node B relays packets received from nodes S_1 and S_2 subject to a strict delay constraint. Assuming the nodes use orthogonal channels, if the transmission rates R_1, R_2 are bounded, then the rates of packets that can be relayed successfully is given by a pentagon (solid line in Fig. 1). Rates in this region are achieved if the relay transmits every received packet after a small processing delay. It is easy to see that such a strategy would result in a high correlation between the source and relay schedules. If, in addition to the networking constraints, the source and relay schedules are forced to be statistically independent, an eavesdropper would not detect correlation across schedules, thus hiding the relaying operation. The delay constraint may, however, result in packet drops or require dummy transmissions thereby reducing the achievable relay rates.

The relaying operation of Figure 1 represents the basic component in wireless networking, and the characterization of the achievable rate region with provable anonymity is one of the contributions of this work. The example highlights that providing anonymity in communication requires a reduction in communication rates. A primary goal of this work is to characterize this trade-off between anonymity and network performance. An analytical approach for the characterization requires a quantifiable notion of anonymity, which we measure using the uncertainty in networking information (active routes in the network) inferable by the adversary. The example discussed suggests a simple technique to provide perfect anonymity by letting all nodes generate statistically independent schedules, but this strategy may not provide scalable performance for large networks. Our goal is to design transmission strategies that sacrifice minimum performance while maintaining a certain level of anonymity.

This work is supported in part by the National Science Foundation under awards CCF-0635070 and CCF-0728872, and the U. S. Army Research Laboratory under the Collaborative Technology Alliance Program DAAD19-01-2-0011. Part of the results in this work were presented in at Allerton 2006 and Allerton 2007.

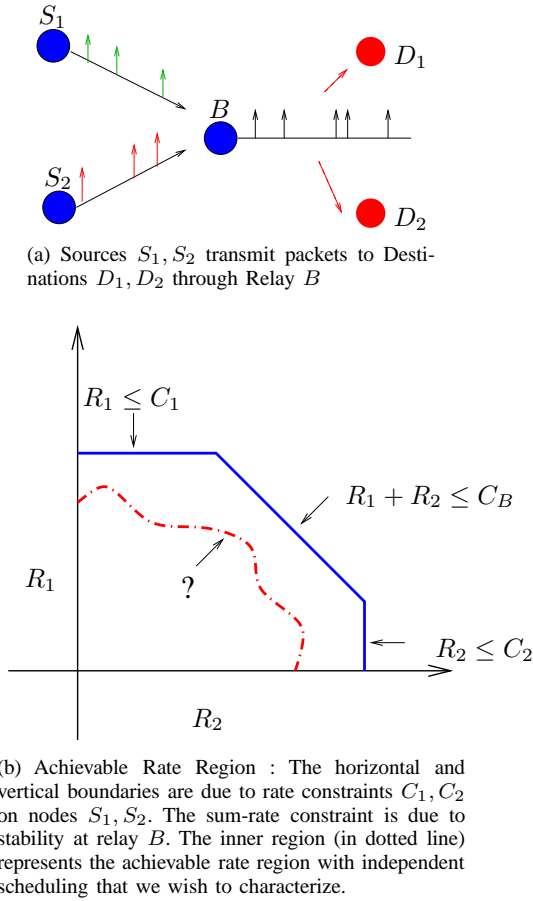


Fig. 1: Two Hop Relay Network

A. Main Contributions

We propose an analytical framework for anonymous scheduling against traffic analysis in wireless networks. In particular, we define a mathematical notion for anonymity of routes, based on Shannon's equivocation [8], when eavesdroppers observe transmission epochs of all nodes in the network. The main results obtained under this model are divided into two segments.

Assuming maximum anonymity requirement, we design scheduling and relaying strategies for a two hop multiple source single relay system (see Fig. 4) when nodes use orthogonal transmitter directed signaling. In particular, when the transmission schedules of nodes are independent Poisson processes, we characterize the achievable rate region analytically. Although independent Poisson scheduling may not be optimal for a strict delay constraint on the relay, we show that, under certain physical layer conditions, the achievable relay rates are optimal for an average delay constraint.

For a general multihop network, we propose a randomized scheduling strategy for any given level of anonymity α , and utilizing the results of the two hop system, characterize the achievable sum-rate of data flows as a function of α . Our key result in this framework shows the equivalence between the sum-rate anonymity tradeoff and information theoretic rate-distortion.

The connection between rate distortion and anonymous

networking is not tied to our strategy and can be explained using a general intuition. The objective of the rate-distortion problem is to generate fewest number of codewords for a set of source sequences, such that the corresponding reconstruction sequences satisfy a specified distortion constraint. The idea is to divide the set of source sequences into fewest number of bins such that the distortion between each sequence in a bin and the reconstruction sequence is less than the specified constraint. Alternatively, fixing the code rate fixes the total number of bins. Then, the sequences are placed optimally within each bin such that the corresponding reconstruction sequences minimize the expected distortion.

In the anonymous networking setup, let the set of active routes at any given time be referred to as a network session. The key idea is to divide the set of all possible network sessions into bins such that, for each bin, there exists a scheduling strategy that would make the sessions within that bin indistinguishable to an eavesdropper. The level of anonymity required determines the number of bins, and the optimal scheduling strategy plays the role of the reconstruction sequence by minimizing the performance loss across sessions within the bin.

B. Related Work

Although prevention of traffic analysis is a classical problem, a dominant portion of prior research has centered around Internet applications. In that regard, an important countermeasure was provided by Chaum through the concept of the traffic Mix [9]. A Mix node uses re-encryption and packet padding to prevent correlation based on contents or lengths across packets. Further, by batching and reordering packets, the Mix provides anonymity of source-destination pairs. Subsequent improvements in the anonymity provided by the Mix included random delaying (Stop-and-Go Mixes [10]) and introducing dummy packets (ISDN Mixes [11]). The concept of Mixes was successfully used in designing remailer and proxy systems [12], [13], [14] for the Internet.

Although Mixes provide an ideal solution for many Internet applications, when strict constraints on delay or buffer size are imposed, it was shown [15] that a Mix no longer provided anonymity to long streams of traffic. An alternative approach, designed primarily for multihop wireless networks is that of deterministic scheduling [16]. In [16], the authors propose a fixed periodic schedule for the entire network, wherein every node adhered to the schedule by transmitting dummy packets whenever actual data was not present. Although the idea of fixed scheduling can be adapted to handle delay constraints, constant transmission of dummy packets is inefficient and furthermore, the centralized synchronous implementation is impractical for ad hoc wireless networks.

A key component of our approach is the analytical model for anonymity of routes. In mix networks, anonymity has been measured using the size or entropy of the anonymity set (set of possible source-destination pairs) of an observed packet. In the context of this work, the use of anonymity sets has two disadvantages. First, hiding source-destination pairs alone may not be sufficient, the direction of data flow could also

reveal critical information. Second, the measure of anonymity needs to cater to streams of packets rather than a single packet [15]. Our metric for anonymity is based on the information theoretic notion of equivocation, proposed by Shannon [8]. Previous applications of equivocation measured the secrecy of transmitted data on point-to-point channels [17], [18], whereas we use equivocation to measure the secrecy of routes in a network.

Prevention of traffic analysis can also be viewed as the complementary problem to intrusion detection [19], which is another important area in network security. Some of the techniques we use to design anonymous relaying strategies are motivated by prior work on stepping stone detection [20].

II. ANALYTICAL MODEL

The main problem addressed in this paper is to design transmission and relaying strategies that are resilient to traffic analysis and use them to characterize the relationship between achievable network performance and the level of anonymity. We consider a specific category of delay sensitive traffic and measure the network performance using achievable packet relay rates from source to destination.

A. Notation

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a directed graph, where \mathcal{V} is the set of nodes in the network and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of directed links. If (A, B) is an element of \mathcal{E} , then node B can receive transmissions from node A . A sequence of nodes $P = (V_1, \dots, V_n) \in \mathcal{V}^*$ is a *valid path*^{*} in \mathcal{G} if $(V_i, V_{i+1}) \in \mathcal{E}, \forall i < n$. The set of all possible paths in \mathcal{G} is denoted by $\mathcal{P}(\mathcal{G})$.

We assume that during any network observation by the eavesdropper, a subset of nodes communicate using a fixed set of paths. This set of paths $\mathbf{S} \in 2^{\mathcal{P}(\mathcal{G})}$ is referred to as a network *session*. The information that we wish to hide from the eavesdropper is the network session \mathbf{S} . We model \mathbf{S} as an i.i.d. random variable with a probability mass function $\{p(\mathbf{s}) : \mathbf{s} \in 2^{\mathcal{P}(\mathcal{G})}\}$. Therefore, the set of all possible sessions is given by

$$\mathcal{S} = \{\mathbf{s} \in 2^{\mathcal{P}(\mathcal{G})} : p(\mathbf{s}) > 0\}.$$

The prior information $p(\mathbf{S})$ on sessions can be obtained using the topology and applications of the particular network, and is also available to the eavesdropper.

For example, in a simple network \mathcal{G}_1 as shown in Figure 2, let S_1, S_2 be the only allowed sources and D_1, D_2 the allowed destinations. Further, let the sources always communicate with distinct destinations. For such a network, $\mathcal{P}(\mathcal{G}_1)$, the set of all possible paths, is given by

$$\mathcal{P}(\mathcal{G}_1) = \{ (S_1, B), (S_1, B, D_1), (S_1, B, D_2), (S_2, B), (S_2, B, D_1), (S_2, B, D_2), (B, D_1), (B, D_2) \}.$$

Due to the restriction on distinct destinations, the set of valid sessions \mathcal{S} contains only two sessions:

$$\mathcal{S} = \{ \{ (S_1, B, D_1), (S_2, B, D_2) \}, \{ (S_1, B, D_2), (S_2, B, D_1) \} \}.$$

^{*}The notation \mathcal{V}^* refers to $\bigcup_i \mathcal{V}^i$.

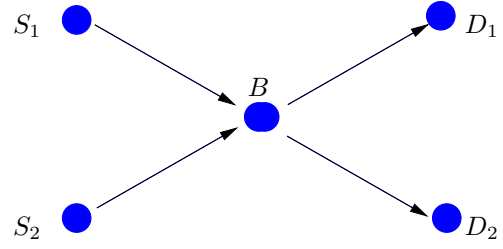


Fig. 2: Two Node Switching Network: $\mathcal{G}_1 = (\mathcal{V}, \mathcal{E})$, $\mathcal{V}_1 = \{S_1, S_2, B, D_1, D_2\}$, $\mathcal{E}_1 = \{(S_1, B), (S_2, B), (B, D_1), (B, D_2)\}$.

Transmission Schedules The eavesdropper's observation consists of the packet transmission epochs in a session. Since it is not possible to determine the location of the eavesdropper(s), we assume that all transmissions are being monitored. Although the packets are encrypted, depending on the physical layer model, it may be possible for an eavesdropper to infer partial information about sender-receiver nodes of packets by merely detecting a transmission. We consider one such physical layer model known as a transmitter directed signaling model.

Transmitter Directed Signaling: All packets transmitted by a particular node are modulated using the same spreading sequence, and each transmitting node is associated with a unique orthogonal spreading sequence. Under this transmission scheme, an eavesdropper would be able to “tune” his detector to a particular spreading sequence and detect the transmission times of packets sent by the corresponding node. Although he knows the transmitting node of each packet, we assume that headers are encrypted, so he would not know the intended recipient of any packet. Therefore, in a route involving multiple nodes, even when all transmission schedules are correlated, it is not possible for an eavesdropper to ascertain the final destination node.

Eavesdropper Observation Let \mathcal{Y}_A represent the schedule of packets transmitted by node A . The schedule \mathcal{Y}_A is a point process,

$$\mathcal{Y}_A = \{Y_A(1), Y_A(2), \dots\},$$

where $Y_A(i)$ represents the transmission epoch of the i^{th} packet by node A . The eavesdropper detects packet transmission epochs which, by virtue of unique orthogonal codes, would provide him the identity of the transmitting node. Since we assume all nodes are monitored, the eavesdropper's complete observation is given by $\mathcal{Y} = \{\mathcal{Y}_A : A \in \mathcal{V}\}$.

Note that, while \mathcal{Y} represents the schedules of packet transmissions detected by eavesdroppers, it does not specify which packets are relayed from sources to destinations in a session. In fact, some of the epochs in \mathcal{Y} could represent dummy transmissions by nodes.

B. Anonymity Measure

We model \mathcal{Y} as a random sequence of epochs with conditional distribution $q(\mathcal{Y}|\mathbf{S})$. The idea is to design $q(\mathcal{Y}|\mathbf{S})$ such that eavesdroppers obtain minimum information about

the session \mathbf{S} by observing \mathcal{Y} . Based on the information we wish to hide (\mathbf{S}) and the observation of the eavesdropper (\mathcal{Y}), we use equivocation [8] to define the analytical measure of anonymity.

Definition 1: A distribution $q(\mathcal{Y}|\mathbf{S})$ is defined to have anonymity α if

$$\frac{H(\mathbf{S}|\mathcal{Y})}{H(\mathbf{S})} \geq \alpha.$$

When $\alpha = 1$, the distribution $q(\mathcal{Y}|\mathbf{S})$ is defined to have *perfect anonymity*. For a distribution with perfect anonymity, given the observed schedules, the eavesdropper gains no additional information (than the prior $p(\mathbf{S})$) about the routes. In other words,

$$H(\mathbf{S}|\mathcal{Y}) = H(\mathbf{S}).$$

For a general α , a physical interpretation of anonymity can be obtained using Fano's Inequality [21]: Let the error probability of the eavesdropper in decoding the session \mathbf{S} be P_e . Then,

$$P_e \geq \frac{H(\mathbf{S}|\mathcal{Y}) - 1}{\log |\mathcal{S}|} \geq \frac{\alpha H(\mathbf{S}) - 1}{\log |\mathcal{S}|}.$$

Furthermore, if \mathcal{S} is a large set with uniform prior $\{p(\mathbf{s}) = \frac{1}{|\mathcal{S}|}, \forall \mathbf{s}\}$, then $P_e \geq \alpha$. In other words, the anonymity bounds the minimum probability of error incurred by the eavesdropper in decoding \mathbf{S} .

This notion of anonymity that we consider is different from previous definitions [22], [10], which were primarily used to hide the source-destination pair of each individual packet. To the best of our knowledge, this is the first definition of anonymity that deals with multihop routes and considers timing information in long streams of transmitted packets.

C. Network Constraints and Throughput

The key challenge in designing the schedule distribution $q(\mathcal{Y}|\mathbf{S})$ with provable anonymity is to sacrifice minimum performance under the networking constraints. In this work, we measure performance using the achievable rates of packets relayed from sources to destinations subject to constraints on medium access and latency, which are described as follows.

Medium Access Constraints Wireless networks, due to restrictions on shared bandwidth and transmission power, pose constraints on rates of packets transmitted and received. We consider long streams of packet transmissions, and measure the rate of packets transmitted using an asymptotic measure:

$$T_A = \lim_{n \rightarrow \infty} \frac{n}{Y_A(n)}, \quad (1)$$

where T_A denotes the rate of packets transmitted by a node A . Since each transmitting node is associated with an orthogonal spreading sequence, the constraint on each point process in \mathcal{Y} is independent. Specifically, the transmission rate T_A of a node A is bounded by a constant C_A , which depends on the characteristics of the medium and the transmission capability of node A . As long as $T_A \leq C_A$, successful reception is guaranteed at the intended receiver.

We assume that the network operates in full duplex mode, where every node can transmit and receive packets simultaneously as long as all transmission rates are within the specified bounds. In other words, a set of schedules \mathcal{Y} is a *valid network schedule* if and only if $T_A \leq C_A$ for every node A .

Latency Constraint: We consider a strict delay constraint on the packets, where the packet delay at each intermediate relay in a route is bounded by Δ . In general, each relay is allowed to reencrypt packets, reorder arrived packets and transmit dummy packets. However, each received data packet at a relay is required to be forwarded within Δ time units of arrival, or otherwise, dropped. Such a strict delay constraint would apply in practice to time sensitive applications such as target tracking in sensor networks or streaming media in peer to peer networks. In general, a strict delay constraint would prevent congestions in the network and ensure stability, albeit at the cost of dropped packets.

Note that the schedules in \mathcal{Y} only specify when packets are transmitted by each node, and do not indicate which packets actually travel from source to destination on each route of a session. For every schedule, we therefore need to specify a relaying strategy, represented by \mathcal{Z} , which is a set of subsequences of \mathcal{Y} . The subsequences represent the transmissions epochs of packets that are relayed from sources to destinations and therefore, depend on the routes of the session as well as the delay constraint.

Definition 2: Let a session $\mathbf{S} = (P_1, \dots, P_{|\mathbf{S}|})$, where $P_i = (A(i, 1), \dots, A(i, m(i)))$ is a valid path of length $m(i)$, and $A(i, j) \in \mathcal{V}$ represents the j^{th} node in path P_i of session \mathbf{S} . A set of subsequences $\mathcal{Z} = \{\mathcal{Z}_{i,j} : i \leq |\mathbf{S}|, j < m(i)\}$ of \mathcal{Y} is a *valid relaying strategy* for \mathbf{S} if:

1. $\forall i, j, \mathcal{Z}_{i,j} \subseteq \mathcal{Y}_{A(i,j)}$.
2. For every i, j, n

$$0 \leq Z_{i,j+1}(n) - Z_{i,j}(n) \leq \Delta.$$

3. If $(A(i, j), A(i, j+1)) = (A(l, m), A(l, m+1))$, then $\mathcal{Z}_{i,j} \cap \mathcal{Z}_{l,m} = \emptyset$.

In the above definition, condition 2 ensures that the relayed packets satisfy the delay constraint Δ at every intermediate relay from the sources to the destinations of the session. Condition 3 ensures that, if any pair of nodes is common to multiple routes, the subsequences picked from the transmission schedules are mutually exclusive.

In Section III-C, we also consider a relaxed version of the delay constraint, where the average delay of packets is bounded at each relay. The definition for a relaying strategy with average delay constraint can be obtained by modifying condition 2 of Definition 2 as:

$$\forall i, j, n, Z_{i,j+1}(n) - Z_{i,j}(n) \geq 0, \quad (2)$$

$$\lim_{n \rightarrow \infty} \sum_{m=1}^n \frac{Z_{i,j+1}(m) - Z_{i,j}(m)}{n} \leq \bar{\Delta}. \quad (3)$$

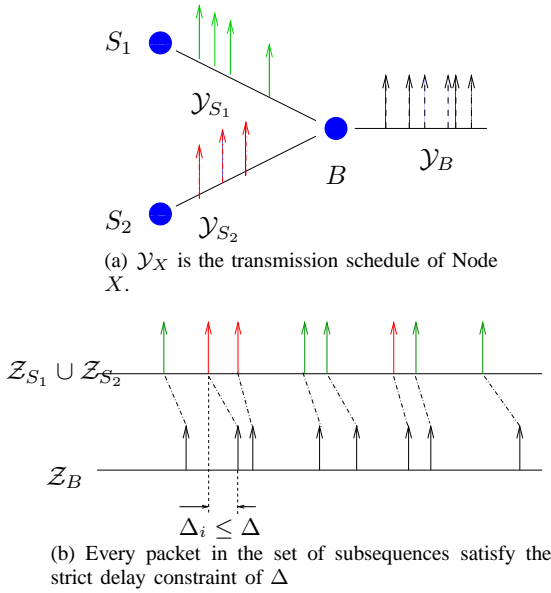


Fig. 3: 2 × 1 Relay with Strict Delay Constraint

D. Performance Metrics

It is possible that the set of subsequences \mathcal{Z} are a strict subset of the transmissions schedule \mathcal{Y} , or in other words, there are epochs in \mathcal{Y} that do not correspond to any relayed packets. Those transmission epochs in \mathcal{Y} that are not present in \mathcal{Z} would either correspond to packets that are dropped eventually, or represent dummy packet transmissions. Therefore, for a session $\mathbf{s} = (P_1, \dots, P_{|\mathbf{s}|})$ and relaying schedule \mathcal{Z} , the rate of packets relayed from source to destination on route P_i is given by:

$$\lambda(\mathcal{Z}, P_i) = \lim_{n \rightarrow \infty} \frac{n}{Z_{i,1}(n)}.$$

Note that, since condition 2 of Definition 2 ensures that all schedules on a route have same length, it is sufficient to use $Z_{i,1}$ to compute rate.

Definition 3: Let the session vector $\mathbf{s} = (P_1, \dots, P_k)$, where $P_i \in V^n$ represents a valid path of data flow. Then, a rate vector $\bar{\lambda}(\mathbf{s}) = (\lambda_1, \dots, \lambda_k)$ is *achievable with strict delay* for session \mathbf{s} if $\exists q(\mathcal{Y}|\mathbf{s})$ with anonymity α such that

1. Every realization of \mathcal{Y} given \mathbf{s} is a valid network schedule.
2. For every realization of \mathcal{Y} , there exists a valid relaying strategy \mathcal{Z} that satisfies

$$\lambda(\mathcal{Z}, P_i) \geq \lambda_i, \forall i. \quad (4)$$

For a large network with several possible session vectors, characterization of the set of rates for each path of each session vector is potentially cumbersome. Furthermore, in order to draw useful inferences on the relationship between anonymity and network performance, it is helpful to have a simpler quantity representing the achievable performance. We, therefore, propose a scalar metric to characterize the performance of large networks, defined by the average sum-rate as follows.

Definition 4: R is defined to be a *weakly achievable sum-rate with anonymity α* if $\exists q(\mathcal{Y}|\mathbf{S})$ with anonymity α such that

1. For every session $\mathbf{s} = \{P_1, \dots, P_{|\mathbf{s}|}\}$, every realization of \mathcal{Y} given \mathbf{s} is a valid network schedule.
2. For every realization of $(\mathbf{S}, \mathcal{Y})$, there exists a valid relaying strategy \mathcal{Z} , and

$$\mathbb{E} \left(\sum_{i=1}^{|\mathbf{S}|} \lambda(\mathcal{Z}, P_i) \right) \geq R, \quad (5)$$

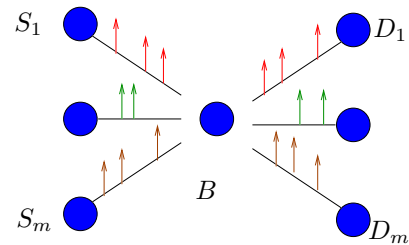
where the expectation is over the joint pdf of \mathcal{Y} and \mathbf{S} .

Note that the rate and sum-rate defined only represent the rate of packets successfully relayed from sources to destinations. Since the relaying strategy could result in packet drops en route to the destinations, the reliability of the achievable rates needs to be proved by specifying packet encoding and decoding techniques. We address this issue using forward error correction in Section III-D.

The fundamental design problem considered in this paper is to characterize the set of achievable rates with anonymity α . Specifically, we derive achievability results for two scenarios: For the two hop network (as shown in Fig. 4), we characterize the set of achievable rate vectors with maximum anonymity ($\alpha = 1$) under both delay constraints. For a general network, we use the results from the two hop network and characterize the weakly achievable sum-rate for a general α .

III. ANONYMOUS MULTIACCESS COMMUNICATION

In this section, we characterize the set of achievable relay rates with maximum anonymity for the two-hop network as shown in Fig. 4. In particular, we provide rate regions for the session vector $\mathbf{s}_m = \{(S_1, B, D_1), (S_2, B, D_2), \dots, (S_m, B, D_m)\}$, i.e. the sources S_1, \dots, S_m transmit packets to destinations D_1, \dots, D_m through relay B .

Fig. 4: Two Hop Network: Source S_i transmits packets to D_i through B

A. Independent Scheduling

In accordance with the definition in Section II-B, scheduling with perfect anonymity corresponds to the independence between session vector \mathbf{S} and the transmission schedules \mathcal{Y} or in other words,

$$H(\mathbf{S}|\mathcal{Y}) = H(\mathbf{S}) \Rightarrow \mathbf{S} \perp \mathcal{Y}.$$

We, therefore, propose an independent scheduling technique, wherein each node in the network generates a random

transmission schedule, statistically independent of the session and the schedules of other nodes in the network. For example, in the network shown in Fig. 4 with $m = 2$

$$q(\mathcal{Y}|\mathbf{S}) = q_1(\mathcal{Y}_{S_1})q_2(\mathcal{Y}_{S_2})q_3(\mathcal{Y}_B),$$

where the distributions q_i do not depend on \mathbf{S} .

Independent scheduling is a particular solution to maintaining anonymity in the two hop setup. An alternative to independent scheduling would be the fixed scheduling as described in [16]. Under that model, all the nodes follow a fixed synchronous schedule irrespective of transmitted data rates or paths of information flow. While the fixed scheduling strategy guarantees maximum anonymity, it would result in a large percentage of dummy packets for low traffic loads. Further, a fixed schedule requires a centralized synchronous implementation, which is impractical in large networks.

The relaying algorithms discussed in this section are not specific to the statistics of the particular transmission processes and some of the optimal properties hold for any pair of point processes. However, for the purpose of analytical characterization of relay rates, we have modeled the transmission schedules to belong to independent Poisson point processes. Poisson processes have typically been used to model the arrival of packets to nodes in a network, due to memoryless interarrival times property. Although Poisson schedules cannot be shown to be optimal under strict delay constraints, under certain conditions on the physical layer, they are shown to be optimal for an average delay constraint. Our relaying algorithms can be used on other point processes, such as Pareto distributed schedules, however the analytical tractability is not guaranteed.

B. Scheduling under Strict Delay

Consider the special case of a single source relay (Fig. 4, $m = 1$). We are interested in the achievable relay rate for the session $s_1 = \{(S_1, B, D_1)\}$. The medium access constraints are specified by the bounds $T_{S_1} \leq C_{S_1}, T_B \leq C_B$ on the transmission rates. If the delay constraint was absent ($\Delta = \infty$), then each received packet can be relayed by B at the next available epoch in its transmission schedule. Since packets can be held for an indefinitely long time, the achievable relay rate would be $\lambda(\mathcal{Z}, (S_1, B, D_1)) = \min\{C_{S_1}, C_B\}$. Note that this is also the maximum possible rate if node B were to relay packets without any anonymity requirement.

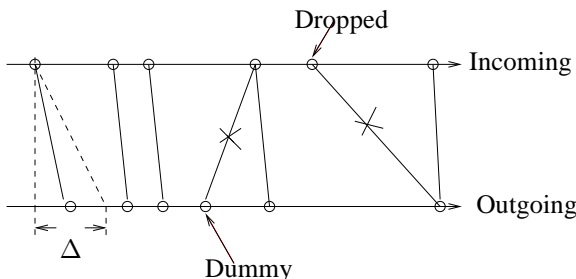


Fig. 5: Bounded Greedy Match: Unmatched packets are dropped, unused epochs have dummy packets

When a strict delay constraint of Δ is imposed, we design the relaying strategy using the *Bounded Greedy Match* (BGM) algorithm proposed in [23] under the context of chaff insertion in stepping stone attacks. The algorithm (Fig. 5) is described in Table I. The basic idea is as follows: When a packet arrives at B , if there exists a departure epoch within Δ of the arrival instant and has not been matched to any previous arrival, it is assigned to the arrived packet. Otherwise, the packet is dropped. If a relay epoch does not have any packet assigned to it, the relay transmits a dummy packet at that epoch.

Let $Y_{S_1}(n), Y_B(n)$ represent the arrival time of the n^{th} packet from S_1 and departure time of n^{th} packet from B .

1. Initialize $i = 1, j = 1$.
2. Let $t = \min\{Y_{S_1}(i), Y_B(j)\}$.
3. If $t = Y_B(j)$, then
 - i. B transmits a dummy packet at time $Y_B(j)$.
 - ii. $j = j + 1$.
- else if $Y_B(j) - Y_{S_1}(i) \leq \Delta$
 - i. B transmits the i^{th} packet from S_1 at $Y_B(j)$.
 - ii. $i = i + 1, j = j + 1$.
- else
 - i. Drop the i^{th} packet that arrived from S_1 .
 - ii. $i = i + 1$.
4. Repeat Step 2,3 until the end of the streams.

TABLE I: Bounded Greedy Match Algorithm

It was shown in [23] that this greedy algorithm resulted in least packet drops. Based on the algorithm, the following theorem characterizes the best achievable relay rate for a pair of independent Poisson processes.

Theorem 1: If the nodes S_1 and B generate independent Poisson transmission schedules, the maximum achievable relay rate from S_1 to D_1 through B is given by $\lambda(\mathcal{Z}, (S_1, B, D_1)) = C_{S_1}(1 - \epsilon(S_1, B))$ where

$$\epsilon(S_1, B) = \begin{cases} \frac{C_B - C_{S_1}}{C_B e^{-\Delta(C_{S_1} - C_B)} - C_{S_1}} & C_{S_1} \neq C_B \\ \frac{1}{1 + C_{S_1} \Delta} & C_{S_1} = C_B \end{cases},$$

$$\triangleq f_e(C_{S_1}, C_B). \quad (6)$$

Proof: Refer to Appendix.

Theorem 1 expresses the maximum achievable rate in terms of the loss function $\epsilon(S_1, B)$ where $\epsilon(S_1, B)$ represents the fraction of packets dropped at relay B . As the delay constraint Δ increases, it is easy to see that the relay rate converges to $\min\{C_{S_1}, C_B\}$ which is the optimal rate under no anonymity requirement. Furthermore, the convergence of the relay rate to the optimal value is exponential in Δ . The value of $\epsilon(S_1, B)$ given in Theorem 1 is obtained when S_1 uses the maximum transmission rate of C_{S_1} for this particular route. In a general network, S_1 could be simultaneously transmitting to another node, in which case, the rate allocated for $\mathcal{Y}_{S_1, B}$ would be strictly less than C_{S_1} . In such a situation, by replacing C_{S_1} in (6) with the allocated rate for the particular flow, we can use Theorem 1 to evaluate the corresponding relay rate.

$m \times 1$ Relay: For the general $m \times 1$ relay as shown in Fig. 4, in the absence of the anonymity constraint, the achievable rate

region can be obtained using the medium access constraints:

$$\bar{\lambda}(s_m) = \{(\lambda_1, \dots, \lambda_m) : \lambda_i \leq C_{S_i} \forall i, \sum_i \lambda_i \leq C_B\}. \quad (7)$$

For a finite delay constraint, a trivial achievable rate region can be obtained if the relay ignores the originating source of the arriving packets. Specifically, the relay uses the BGM algorithm on the joint incoming schedule $\bigcup \mathcal{Y}_{S_i, B}$ and the generated outgoing schedule \mathcal{Y}_B . For this strategy, the single source result in Theorem 1 can be easily extended to characterize an achievable rate region for s_m , which is given in Corollary 1.

Corollary 1: There exists a relaying strategy for a $m \times 1$ relay such that the achievable rates $\bar{\lambda}(s_m) = (\lambda_1, \dots, \lambda_m)$ satisfy $\lambda_i = T_i(1 - \epsilon(S_i, B))$, $\forall i$ where

$$\epsilon(S_i, B) \geq f_e\left(\sum_{j=1}^m T_j, C_B\right), \forall i \quad (8)$$

$$T_i \leq C_{S_i}, \forall i. \quad (9)$$

Prioritized Scheduling Ignoring the source identities and considering the joint stream is strictly sub-optimal. Since the relay observes a distinct stream from each source node (by virtue of transmitter directed signaling), the streams can be prioritized to obtain a larger achievable rate region compared to Corollary 1.

Consider a 2×1 relay and assign the highest priority to S_1 . For every departure epoch in \mathcal{Y}_B , the relay considers all packets that have arrived within Δ time units before that epoch. If some of those packets arrived from S_1 (highest priority), the relay transmits the earliest of those packets at the chosen epoch. If none of the packets arrived from S_1 , then the packet that arrived first (from S_2) is transmitted. Since S_1 is given highest priority, this would provide the maximum rate achievable for the stream from S_1 . The priority algorithm is formally described in Table II.

- | |
|--|
| <ol style="list-style-type: none"> 1. Initialize $i = 1, j = 1, k = 1$. 2. If $Y_B(j) - Z_{S_1, B}(i) \geq \Delta$ <ol style="list-style-type: none"> i. Drop i^{th} packet from S_1. ii. $i = i + 1$. Repeat Step 2. 3. Let $t = \min\{Z_{S_1, B}(i), Y_B(j)\}$. 4. If $t = Y_B(j)$ <ol style="list-style-type: none"> i. Let $t' = \min\{Z_{S_2, B}(j), Y_B(k)\}$. ii. If $t' = Y_B(k)$ then B transmit dummy packet at t'. $k = k + 1$. else if $Z_{S_2, B}(j) \geq Y_B(k) - \Delta$ B transmits j^{th} packet from S_2. $j = j + 1, k = k + 1$. else $j = j + 1$. Repeat Step 4.ii. else B transmits i^{th} packet from S_1. $i = i + 1, k = k + 1$. 5. Repeat Steps 2-4 until end of streams. |
|--|

TABLE II: Priority Mapping Algorithm: S_1 highest priority

Similarly, by interchanging the priorities, we can obtain the maximum rate for the stream from S_2 . It is easy to see that, when none of the sources are given priority, it is equivalent to ignoring the origin of packets (Corollary 1). By time-sharing multiple relaying strategies with different priority

requirements, a piece-wise linear region of achievable rate vectors is obtained, which is characterized in Theorem 2.

Theorem 2: If $\bar{\lambda}(s_2) = (\lambda_1, \lambda_2)$ represents the achievable relay rates for sources S_1 and S_2 through relay B , then

1. (λ_1, λ_2) is achievable if

$$\lambda_1 \leq a_1 \lambda_2 + b_1, \lambda_2 \leq a_2 \lambda_1 + b_2, \lambda_i \leq C_{S_i}(1 - f_e(C_{S_i}, C_B)), \quad (10)$$

where $j \neq i$ and

$$a_i = \frac{C_{S_i}}{C_{S_j}} + \frac{C_B[(1 + \Delta(C_B - C_{S_i} - C_{S_j}) - 1)]}{C_{S_i}(C_B e^{\Delta(C_B - C_{S_i} - C_{S_j})} - C_B)(C_B e^{\Delta(C_B - C_{S_i} - C_{S_j})} - C_{S_i} - C_{S_j})},$$

$$b_i = (C_{S_i} - C_{S_j})a_1 f_e(C_{S_i} + C_{S_j}, C_B).$$

2. (λ_1, λ_2) is not achievable if

$$\sum_i \lambda_i \geq (C_{S_1} + C_{S_2})(1 - f_e(C_{S_1} + C_{S_2}, C_B)), \lambda_i \geq C_{S_i}(1 - f_e(C_{S_i}, C_B)) \quad (13)$$

Proof: Refer to Appendix.

The priority scheduling cannot be proven to obtain the optimal achievable rate region, and so Theorem 2 also provides an outer bound to determine the extent of possible sub-optimality. The outer bound is an upper bound on the sum rate $\lambda_1 + \lambda_2$ that is obtained using the optimality of the BGM algorithm. It can be shown that as $\Delta \rightarrow \infty$, the inner and outer bounds coincide and converge exponentially fast. Although the optimality of the region for Poisson processes is still an open problem, the strategy achieves the maximum possible sum-rate.

The prioritized scheduling can be extended to a general $m \times 1$ relay. Every priority assignment corresponds to an ordering of the sources. When packets from multiple sources contend for a single epoch, the choice of packet to relay is made according to the ordering. Further, by time-sharing strategies for different priority assignments, the complete region can be obtained.

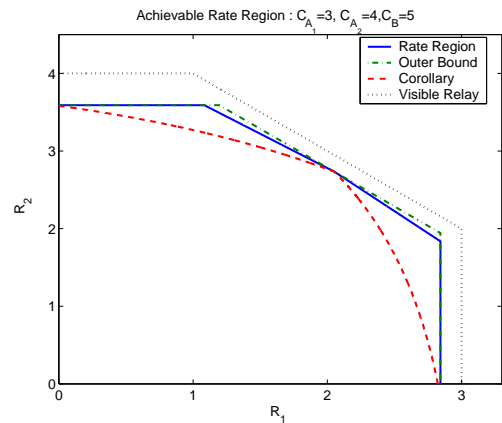


Fig. 6: 2×1 Relay rate region. R_i is the rate $\lambda(Z, (S_i, B, D_i))$. The inner and outer bounds coincide at the maximal sum-rate point.

An example region for the 2×1 relay is shown in Fig. 6. As is evident, the time-sharing strategy results in a piece-wise linear and convex region. The two corner points of the polygon

in the figure which represent the achievable rate-pairs when S_2, S_1 are respectively given full priority, clearly demonstrate the gains due to prioritized scheduling. Even when S_1 is given full priority, the relay rate for S_2 is strictly positive. If no priority is used, however, S_1 can achieve maximum rate only when S_2 does not transmit at all (region of Corollary 1). The maximum priority rate-pairs can also be viewed as the outcome of successive application of the BGM algorithm on the incoming streams from the two sources, with the order of application determined from the priority assignment.

From theorems 1 and 2, it is clear that when C_{S_i}, C_B and Δ are finite, the relay rates are strictly less than the transmission rates, thereby resulting in a non-zero packet drop rate. Therefore, the source needs to employ forward error correction (FEC) in order to deliver information to the destination reliably. It can be shown that for very long streams, the coding does not result in further rate reduction (see Section III-D).

C. Average Delay

In this section, we consider the average delay constraint at a relay, as specified by (2) and (3). It is easy to see that achievable rate regions for an average delay constraint of $\bar{\Delta}$ can be trivially obtained by using the algorithms of Section III-A that assume a strict delay of $\bar{\Delta}$. This trivial strategy, however, can be significantly improved by modifying the algorithms appropriately.

Consider the single source relay. Let $m(\Delta, C_{S_1}, C_B)$ represent the mean packet delay obtained when the BGM algorithm is applied with strict delay constraint Δ . Since we consider infinitely long streams with an asymptotic constraint, we can choose a strict delay constraint Δ^* such that the mean delay $m(\Delta^*, C_{S_1}, B) = \bar{\Delta}$.

Theorem 3: $\lambda(Z, (S_1, B, D_1)) = C_{S_1}(1 - \epsilon(S_1, B))$ is an achievable relay rate for an average delay constraint of $\bar{\Delta}$ if

$$\epsilon(S_1, B) \geq \begin{cases} f_e(\Delta^*, C_{S_1}, C_B) & C_B - C_{S_1} \leq \frac{1}{\bar{\Delta}} \\ 0 & \text{o.w.} \end{cases}$$

and Δ^* is the solution to $m(\Delta^*, C_{S_1}, C_B) = \bar{\Delta}$ where

$$m(\Delta^*, C_{S_1}, C_B) = \frac{1 + e^{\Delta^*(C_{S_1} - C_B)} [\Delta^*(C_{S_1} - C_B) - 1]}{(C_B - C_{S_1}) [1 - e^{\Delta^*(C_{S_1} - C_B)}]}.$$

Proof: Refer to Appendix

For values of $\bar{\Delta}$ close to zero, the strict delay constraint $\Delta^* \approx 2\bar{\Delta}$. Therefore, for very small delays, an average delay constraint does not provide significant improvement in achievable rate compared to a strict delay constraint. However, as $\bar{\Delta}$ increases beyond a certain threshold, the equivalent strict delay Δ^* increases exponentially. In that regime, an achievable rate close to optimal can be obtained even for a bounded $\bar{\Delta}$. Furthermore, as is evident from the Theorem, when $C_B - C_{S_1} \geq \frac{1}{\bar{\Delta}}$, the strategy achieves zero packet loss. In other words, every transmitted packet can be relayed successfully within the (average) delay constraint.

Since we consider long streams, this strategy could potentially be improved by dividing the stream into finite number (N) of segments, and implementing the BGM algorithm with

a different strict delay constraint (Δ_i^*) in each segment (see Fig. 7). The strict delay constraints should be chosen such that the average delay $\frac{\sum_i m(\Delta_i^*, C_{S_1}, B)}{N}$ is less than $\bar{\Delta}$. As the length of the stream increases, each segment i would provide an achievable relay rate $\lambda^i = C_{S_i}(1 - f_e(\Delta_i^*, C_{S_i}, C_B))$ (Theorem 1) and the net achievable rate would be $\frac{\sum_i \lambda^i}{N}$. However, for a pair of Poisson processes, it can be shown that λ^i is a convex function of the strict delay Δ_i^* , and hence, this segmentation does not reduce[†] packet loss for a fixed average delay.

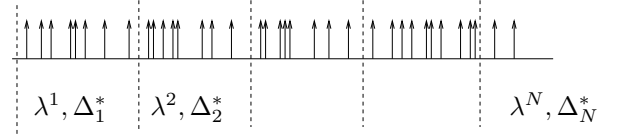


Fig. 7: Delay Segmentation: In each segment of the traffic, a different strict delay Δ_i^* is chosen.

Using the relation between the strict delay and average delay in Theorem 3, the achievable region for the $m \times 1$ relay can also be obtained by appropriately modifying the strict delay constraint in the prioritized scheduling. The condition on transmission rates for which the priority scheduling strategy is optimal for the $m \times 1$ relay case is a straightforward extension of Theorem 3.

Corollary 2: There exists a scheduling strategy for average delay $\bar{\Delta}$ that incurs zero packet loss on all incoming streams, if the medium access constraints satisfy:

$$C_B - \sum_i C_{S_i} \geq \frac{1}{\bar{\Delta}}.$$

From the results presented so far, it is clear that while independent Poisson scheduling generally provides a subset of achievable relay rates for strict delay constraints, under certain conditions on the medium access, it can be optimal for an average delay constraint. An important feature in the algorithms presented is that the relays do not require prior knowledge about transmission schedules of the source nodes. The decision to transmit any packet is based on events occurring between its arrival time and the subsequent departure epoch. This makes it particularly attractive for a decentralized implementation of the scheduling, which is of particular value in adhoc wireless and sensor networks. Note that although the rate expressions derived are for Poisson processes, the algorithms presented are quite general, and can be used on any set of point processes. Furthermore, the optimality of the BGM algorithm also holds for any pair of point processes.

D. Reliability

The independent schedules and relaying algorithms discussed previously result in strictly non-zero packet drop rate for Poisson processes. Further, since the relay nodes generate schedules in a decentralized manner, it is not possible for the source node to know the identities of packets that would be

[†]This convexity may not hold for non-Poisson schedules, in which case, the segmentation could potentially increase the achievable relay rate.

dropped. This implies that the source nodes must employ forward error correction (FEC) techniques to transmit information reliably to the destination. When the traffic is time sensitive such as in media transmission, FEC may not be practical, as it would incur significant coding delay. However, if the strict delay constraint is enforced due to low duty cycles (as in sensor networks) or to maintain stability, it is useful to employ coding to ensure reliability of transmission.

In order to analyze the reliability of packet transmissions, it is necessary to characterize the channel model between a source and destination. For this purpose, if we treat each packet as a binary unit of data, then the packet drops can be equated to a binary erasure channel. Since packets can be appended with indices, the erasure positions would be known at the destination node.

Consider a relay node forwarding packets from a single source. Let $E(i)$ denote the random variable indicating that packet i was successfully relayed when applying the BGM relay algorithm. Then, using Proposition 4 in [24], it can be shown that the relay rate obtained from Theorem 1 can be achieved reliably.

Lemma 1: The capacity C of the erasure channel for a single source relay after applying the BGM algorithm is

$$C = 1 - \limsup_n \frac{1}{n} \sum_{i \leq n} E(i) = 1 - \epsilon(S_1, B),$$

where $\epsilon(S_1, B)$ is given by (6).

Proof: Refer to Appendix.

The achievability of this reliable rate, however, requires coding across a long stream of packets. Since prioritized scheduling is equivalent to successive application of the BGM algorithm, the rate region of Theorem 2 also represent reliable rates. In practice, a packet is not a unit of data and the FEC is different from regular point to point communication channels. Coding for packet recovery in networks has been addressed in literature [25], [26]. In particular, in [25], the authors propose coding schemes, where, for every block of information packets, parity packets are transmitted such that $\forall i$, the i th bit from each packet arranged in sequence forms a codeword from an erasure correcting codebook.

IV. SUM-RATE SECRECY REGION

The achievability results presented in the previous section can be viewed as the basic building blocks for hiding routes in a network. While the independent scheduling idea can be directly extended to multihop routes, characterizing rate regions for large networks is cumbersome and not practical. Furthermore, Theorem 2 in [27] shows that under certain conditions, for an n -hop path with independent Poisson schedules, the maximum rate of packets that can be relayed to the destination with strict delay constraint decays exponentially as n increases. Therefore, instead of directly extending the idea, we propose to utilize independent scheduling at selected portions of the network depending on the required level of anonymity α .

As an example, consider the switching network shown in Fig. 8. During any network session, each source S_i picks a

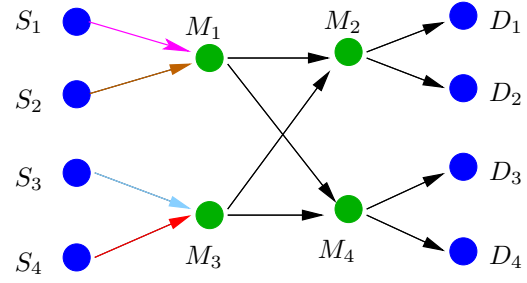


Fig. 8: Switching Network: Sources $\{S_i\}$ transmit packets to destinations $\{D_i\}$ through relays $\{M_i\}$.

distinct destination D_j . It is easy to see that given the S_i, D_j pairings, there is a unique set of paths in the session S . If no anonymity is required, each relay would transmit a received packet after a negligible processing delay, thereby incurring no packet drops. Assuming each node has a transmission rate of C , the average sum-rate achievable would be $2C$ (min-cut would be out of M_1, M_3). Since the schedules of all the relays are dependent on the arrival processes, the eavesdropper would be able to detect the relaying operation of the nodes M_1, \dots, M_4 . However, since nodes utilize transmitter directed signaling with encrypted headers, the eavesdropper would not be able to ascertain the final destination nodes of any path. In this case, it can be shown that the anonymity level $\frac{H(S|Y)}{H(S)} = .436$.

On the other hand, complete independent scheduling would imply that the relays M_1, \dots, M_4 generate statistically independent schedules. Such a strategy would provide maximum anonymity $\alpha = 1$, but result in a reduced achievable sum-rate given by $2C(1 - \epsilon_1)(1 - \epsilon_2)$, where ϵ_1, ϵ_2 are packet losses incurred at relays M_1, M_3 and M_2, M_4 respectively.

Suppose, only M_1, M_3 were to generate independent schedules, while M_2, M_4 relayed packets immediately, the eavesdropper would be able to observe a portion of the paths. In that case, it can be shown that the anonymity level $\frac{H(S|Y)}{H(S)} = .65$ (refer to Appendix for details). However since only one relay in each path drops packets, the achievable sum-rate, however, increases to $2C(1 - \epsilon_1)$.

This simple example illustrates the trade-off between achievable network performance and the level of anonymity. In the remainder of this section, we shall formalize these ideas, describe a randomized relaying strategy and provide an analytical characterization of the achievable sum-rate as a function of anonymity.

A. Relay Categories

As suggested in the example, the key idea we exploit is to divide the set of relays according to their scheduling strategies. Specifically, we categorize the relays into two types: *covert* and *visible* relays.

Covert Relays: A relay M is *covert*, if it generates a transmission schedule statistically independent of the schedules of all nodes occurring previously in the paths that contain M . For example, if only path $P = \{A_1, \dots, A_k, M, A_{k+1}, \dots\}$ contains M , then M is covert if its transmission schedule is

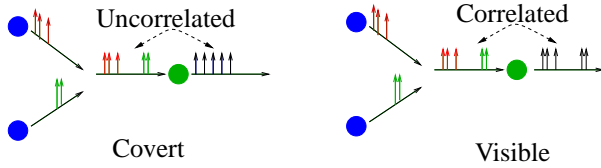


Fig. 9: Visible and Covert Relaying.

independent of schedules of A_1, \dots, A_k . Further, if M relays packets from k nodes, then it uses the BGM algorithm on the joint incoming packet stream to optimally match the departure epochs. Since our criterion is to maximize sum-rate, the nodes are given equal priority which is the sum-rate optimal strategy (Theorem 2).

Visible Relays: A visible relay M generates its schedule based on the schedules of nodes transmitting packets to M . For every received packet, the relay schedules an epoch after a processing delay (negligible compared to Δ). It is evident that a relay operating under this highly correlated schedule would be easily detected by an eavesdropper. It is important to note that, although some received packets from the transmitting node may be dummy packets, these are also relayed by a visible node. The reason is that, if dummy packets that were generated due to independent scheduling at a previous node were to be dropped by the visible relay, then the new stream would no longer be independent from the node two hops earlier (see Fig. 10). We assume that for visible relays, the eavesdropper makes a perfect detection of the relaying operation.

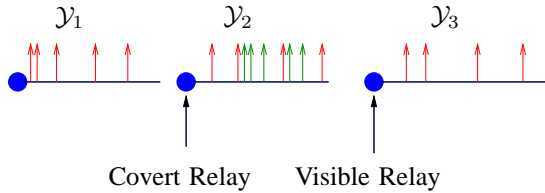


Fig. 10: Relaying Dummy Packets: \mathcal{Y}_1 and \mathcal{Y}_2 are statistically independent. If the dummy packets (represented in green) are not relayed, the processes \mathcal{Y}_1 and \mathcal{Y}_3 will be dependent.

By appropriately selecting which relays should be covert in a session, we can guarantee the required level of anonymity. A trivial strategy would be to let all nodes act as covert relays in a session. However, since the independent schedules would result in packet loss at every covert relay, network throughput would be reduced significantly. It is, therefore, necessary to pick the covert relays optimally so that anonymity is guaranteed with minimum loss in throughput.

We assume the transmission times of packets by each source node in a session are generated according to an independent Poisson process. To maintain uniformity in traffic schedule patterns, we let the covert relays also generate independent Poisson processes. Given a session \mathbf{S} , let \mathbf{B} represent the set of relay nodes that are chosen to be covert. Given \mathbf{S}, \mathbf{B} , using the relaying algorithms discussed in the previous section, the

schedules \mathcal{Y} and the relayed subsequences \mathcal{Z} can be generated for all nodes in the network.

B. Eavesdropper Observation

We assume that when a relay is visible, the eavesdropper perfectly correlates the schedules transmitted by a preceding node and the relay. As a result, depending on the set of visible relays, the eavesdropper makes a partial detection on the paths of a session. We denote this partial observation as a set of paths, $\hat{\mathbf{S}} \in 2^{\mathcal{P}(\mathcal{G})}$. Given the observation $\hat{\mathbf{S}}$, the eavesdropper would try and infer the actual session \mathbf{S} . The partial observation $\hat{\mathbf{S}}$ can be expressed as a function of the actual session \mathbf{S} and the set of covert relays \mathbf{B} .

We define function $t : 2^{\mathcal{P}(\mathcal{G})} \times \mathcal{V} \rightarrow 2^{\mathcal{P}(\mathcal{G})}$ to characterize the eavesdropper's observation when at most one relay is covert. For a set of paths \mathbf{P} , $t(\mathbf{P}, B)$ contains the observed paths when only node B is covert. If $B = \phi$, then $t(\mathbf{P}, \phi)$ is obtained by removing the destination nodes from every path in \mathbf{P} . This is because, even if all relays are visible, transmitter directed signaling ensures that it is not possible to detect the final destination in any route. If $B \neq \phi$, then a path $P \in \mathcal{P}(\mathcal{G})$ belongs to $t(\mathbf{P}, B)$ if and only if it satisfies one of the following conditions:

1. $\exists P' = (A_1, \dots, A_k, B, A_{k+1}, \dots, A_n) \in \mathbf{P}$, such that $P = (A_1, \dots, A_k)$ or $P = (B, A_{k+1}, \dots, A_n)$.
2. $P \in \mathbf{P}$ and $B \notin P$.

Condition 1 states that, when a path in \mathbf{P} contains a covert relay, the eavesdropper would observe two different paths, one terminating before B and the other originating from node B . Condition 2 states that a path that does not contain a covert relay is fully observed. When a subset $\mathbf{B} = (B_1, \dots, B_m) \subset \mathcal{V}$ of relays are covert, then $\hat{\mathbf{S}}$ can be obtained by repeated application of t :

$$\hat{\mathbf{S}} = t(\dots(t(t(\mathbf{S}, \phi), B_1) \dots), B_m) \triangleq \mathbf{T}(\mathbf{S}, \mathbf{B}). \quad (14)$$

It can be shown that the set $\hat{\mathbf{S}}$ in the above equation, represents the eavesdropper's sufficient statistic (part of the proof of Theorem 4).

C. Throughput Function

In order to design the optimal selection strategy, we first characterize the loss in sum-rate when a deterministic set of relays are covert in a session. The relaying strategies in Section III-A were designed to minimize the packet loss at a single covert relay. Extending those results to multihop routes, we can characterize the loss in sum-rate of each session \mathbf{S} , when a subset of relays \mathbf{B} are covert.

If we ignore the anonymity requirement, the best throughput in the network is achieved when all relays are visible. Each session \mathbf{S} corresponds to a maximum achievable sum-rate obtained using the max-flow that satisfies medium access constraints. Specifically, let $\bar{\lambda}^v(\mathbf{S}) = (\lambda_1^v, \dots, \lambda_{|\mathbf{S}|}^v)$ represent the vector of achievable relay rates for the paths in session \mathbf{S} with no covert relays, and $\Lambda^v(\mathbf{S})$ be the maximum achievable sum-rate.

If $\mathbf{S} = (P_1, \dots, P_{|\mathbf{S}|})$, then, using the forwarding strategy for visible relays, the maximum achievable sum-rate is the solution to:

$$\Lambda^v(\mathbf{S}) = \max(\lambda_1^v + \dots + \lambda_k^v), \quad (15)$$

$$\sum_{i: B \in P_i} \lambda_i^v \leq C_B, \quad \forall B \in V. \quad (16)$$

Therefore our performance metric when anonymity $\alpha = 0$ is the maximum expected sum-rate given by,

$$R(\alpha = 0) = \mathbb{E}(\Lambda^v(\mathbf{S})),$$

where the expectation is over the prior $p(\mathbf{S})$. Although in practice, the actual rates of flows are dependent on the nature of data and network application, the maximum sum-rate is a metric that represents the fundamental limits of achievable performance.

When a subset of relays are covert, the achievable sum-rate in each session is reduced depending on the fraction of packets dropped at each covert relay. The net relay rate for each path is obtained by multiplying the fraction of packets that are relayed at every covert relay in that path.

Specifically, let $\lambda^c(\mathbf{S}, \mathbf{B}) = (\lambda_1^c, \dots, \lambda_{|\mathbf{S}|}^c)$ represent the achievable relay rates from sources to destinations for a session $\mathbf{S} = (P_1, \dots, P_{|\mathbf{S}|})$, when nodes in \mathbf{B} are covert, and let $\Lambda^c(\mathbf{S}, \mathbf{B}) \triangleq \sum_{i=1}^{|\mathbf{S}|} \lambda_i^c$ be the achievable sum-rate. If $A(i, j)$ represents the j^{th} node in path P_i , then

$$\lambda_i^c = \lambda_i^v \prod_{j: A(i, j) \in \mathbf{B} \cap P_i} (1 - \epsilon_i(A(i, j-1), A(i, j))) \quad (17)$$

where $\epsilon_i(A, B)$ represents the fraction of packets transmitted by node A on path P_i , that are dropped by covert relay B . Note that Theorems 1 and 2 provide the closed form expression for $\epsilon_i(A, B)$, if B is the first covert relay in the path i . Since the departure epochs of data packets from a covert relay do not constitute a Poisson process, the expression cannot be applied to subsequent covert relays. The analytical characterization of multiple covert relays is generally cumbersome, but can be obtained numerically.

Although the solution of the optimization in ((15),(16)) specifies a set of transmission rates for the nodes, we know from Theorems 1 and 2 that, increasing the transmission rates of nodes results in lower packet losses for statistically independent schedules. Therefore, if the relay immediately following a source node is covert, the source node could transmit at the maximum rate possible to minimize packet losses. In other words, if A is a source node, then $T_A = \sum_{i: A \in P_i} \lambda_i^v$ can be increased to C_A . Since only the source is allowed to perform forward error correction, it does not help to increase transmission rates of subsequent relays (as we would only get additional dummy packets).

V. PERFORMANCE CHARACTERIZATION

With the eavesdropper observation of (14) and throughput characterization in (17), we now have all the elements required to maximize throughput with anonymity α . Prior to describing the general randomized strategy, to ease understanding, we first discuss a simple deterministic strategy to obtain a

smaller region of achievable sum-rate anonymity pairs. Then, expanding on that idea, we provide the generalized strategy to characterize the sum-rate anonymity region.

Deterministic Covert Scheduling: A direct optimization of (17) provides a deterministic strategy to characterize achievable sum-rates under anonymity constraints. Specifically, a subset \mathbf{B} of relays is chosen to remain covert for all sessions, such that the sum-rate is maximized without violating the anonymity requirement.

Theorem 4: A sum-rate R is achievable with anonymity α if

$$R \leq \max_{\mathbf{B}: H(\mathbf{S}|\hat{\mathbf{S}}) \geq \alpha} \mathbb{E}[\Lambda^c(\mathbf{S}, \mathbf{B})],$$

where $\hat{\mathbf{S}} = \mathbf{T}(\mathbf{S}, \mathbf{B})$.

Proof: Refer to Appendix

Depending on the level of anonymity required, the strategy picks one subset of nodes that are always covert (for all sessions). Since the number of possible subsets is finite, the achievable sum-rate anonymity region would be constant within intervals of α , with sudden jumps corresponding to a change in the optimal subset (see example in Section VI).

The above theorem provides one set of achievable sum-rates as a function of anonymity α . As mentioned in Section II-B, equivocation is an average metric. It gives a lower bound on the *average* probability of error for the adversary. Furthermore, the performance is also measured by an average sum-rate metric. Therefore, by time-sharing multiple strategies, it is possible to obtain a convex region without violating the anonymity constraint.

For example, let two subsets of covert relays \mathbf{B}_1 and \mathbf{B}_2 correspond to achievable sum-rate anonymity pairs R_1, α_1 and R_2, α_2 . At the beginning of every session, one of the subsets $\mathbf{B}_1, \mathbf{B}_2$ are chosen with probability $\frac{1}{2}$. Then, it is possible to obtain an achievable sum-rate anonymity pair $(\frac{R_1+R_2}{2}, \frac{\alpha_1+\alpha_2}{2})$. In general, any convex combination of sum-rate anonymity pairs is achievable by time-sharing.

Corollary 3: Let

$$\mathcal{R}^{\text{det}} = \{(R, \alpha) : R \text{ is an achievable sum-rate with anonymity } \alpha\}.$$

Then, every $(R, \alpha) \in \text{convex-hull}(\mathcal{R}^{\text{det}})$ is achievable.

Randomized Covert Scheduling: The drawback in the strategies discussed above is that the subset \mathbf{B} is chosen independent of the session \mathbf{S} . The generalized strategy is to choose the set of covert relays as a random function of the session \mathbf{S} . We model the set of covert relays \mathbf{B} as a random variable with a conditional probability mass function $\{q(\mathbf{B}|\mathbf{S}) : \mathbf{B} \in 2^V\}$. The goal is to optimize the conditional p.m.f $\{q(\mathbf{B}|\mathbf{S})\}$ so that achievable sum-rate is maximized for a given level of anonymity α . Obtaining the best distribution could typically be done using a brute force optimization over a large dimensional simplex, which is computationally intensive, and impractical for large networks. However, the following result proves the duality of this problem to information theoretic rate-distortion, which can then be used to efficiently obtain the optimal strategy and characterize the optimal sum-rate $R(\alpha)$.

Theorem 5: Let $d : 2^{\mathcal{P}} \times 2^{\mathcal{P}} \rightarrow \mathcal{R}$ s.t.

$$d(\mathbf{S}, \hat{\mathbf{S}}) = \begin{cases} \Lambda^v(\mathbf{S}) - \Lambda^c(\mathbf{S}, \mathbf{B}) & \exists \mathbf{B} \text{ s.t. } \hat{\mathbf{S}} = T(\mathbf{S}, \mathbf{B}) \\ \infty & \text{o.w.} \end{cases} \quad (18)$$

Then, a sum-rate $R(\alpha)$ is achievable with anonymity α if

$$R(0) - R(\alpha) \geq D(H(\mathbf{S})(1 - \alpha)),$$

where $D(r)$ is the *Distortion-Rate* function defined as

$$D(r) = \min_{q(\hat{\mathbf{S}}|\mathbf{S}): I(\mathbf{S}; \hat{\mathbf{S}}) \leq r} \mathbb{E}(d(\mathbf{S}, \hat{\mathbf{S}})). \quad (19)$$

Proof: Refer to Appendix.

The above theorem provides $R(\alpha)$ using the single letter characterization of a rate-distortion function. The loss function $d(\mathbf{S}, \hat{\mathbf{S}})$ represents the reduction in sum-rate due to covert relaying. Although the loss function parameters do not explicitly include the set of covert relays \mathbf{B} , it can be shown that given $\mathbf{S}, \hat{\mathbf{S}}$, the set of covert relays \mathbf{B} is unique (see proof of Theorem 4). Therefore, the distribution $q(\mathbf{B}|\mathbf{S})$ to chose covert relays is equivalent to the distortion minimizing distribution in (19). As a result, the Blahut-Arimoto algorithm [28] provides an efficient iterative technique to obtain $q(\mathbf{B}|\mathbf{S})$ and the achievable sum-rate $R(\alpha)$. Note that the anonymity α is guaranteed assuming that the eavesdropper is aware of the network topology, the session prior distribution $p(\mathbf{S})$ and the optimal strategy $q(\mathbf{B}|\mathbf{S})$ of choosing covert relays.

A. Discussion

The equivalence between anonymous networking and rate distortion is not tied to our strategy of choosing covert relays, as explained in Section I-A. In our model, the level of anonymity α directly corresponds to the rate of compression and the performance loss function plays the role of distortion. Therefore, obtaining the optimal rate-distortion function is equivalent to obtaining the throughput anonymity relation.

We believe that the consequences of this duality extend beyond the characterization of the tradeoff between anonymity and throughput. Rate distortion is a field that has been studied for many decades [21], and the numerous models and techniques developed therein could serve to design strategies for anonymous networking. For example, in our setup, the Blahut-Arimoto algorithm provides an efficient iterative technique to obtain the optimal distribution of covert relays in a session.

In our current setup, we have considered independent sessions of observation, which may not apply to the scenario where an eavesdropper monitors the network for long periods of time. In that case, we would need a stochastic model to account for session changes, depending on when nodes start or stop communication. Based on the duality we believe that, if we adopt a Markovian model for the session evolution, then techniques in causal source coding [29] would provide possible solutions.

We currently model the entire session as a single entity (the variable \mathbf{S}) which may not be practical to analyze in a large scale network. This model should be broken down to protecting each route independently, depending on the level of anonymity required by that particular route. One approach

towards such a model would be to express the set of routes as sequence of links, rather than a single session variable. Each session would then be correspond to a source sequence, and the distortion measure would depend on the relative levels of anonymity required by routes. The challenge in developing such a model, however, is to account for eavesdroppers correlating schedules across multiple hops.

VI. EXAMPLE

Consider the switching example given in the beginning of Section IV (Fig. 8). During any network session, each source S_i picks a distinct destination D_i . The set of sessions \mathcal{S} , contains 24 elements which are assumed equiprobable. For this example, Fig. 11 plots the sum-rate anonymity region for the deterministic and probabilistic strategies discussed previously.

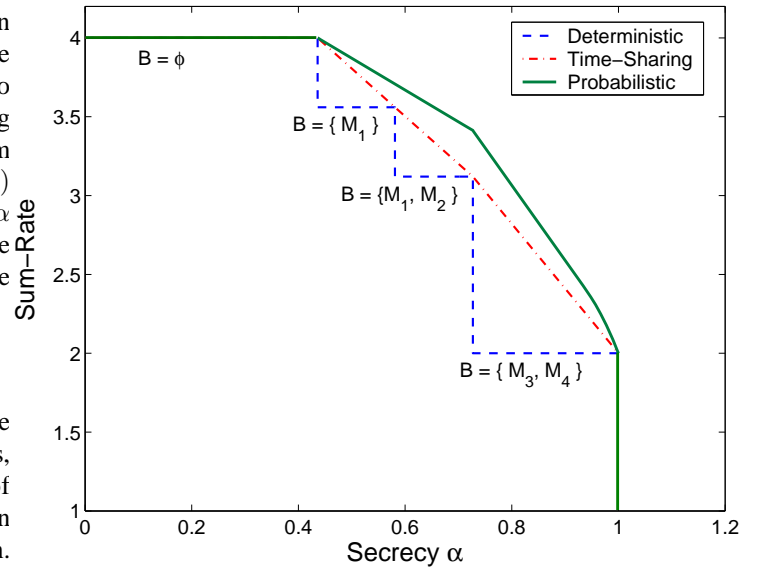


Fig. 11: Sum-Rate Anonymity Region for 4×4 switching network with $C = 2$.

The sum-rate anonymity relationship is convex as seen in the figure. This is because the performance metrics, namely anonymity and throughput, are average quantities, which allows time-sharing to convexify any set of achievable rates. The figure clearly demonstrates the performance improvement due to the randomized covert scheduling. As can be seen, when all relays are visible, the maximum sum-rate $2C$ is achieved with a strictly positive secrecy level. This is because, given the transmission stream from relay M_2 (or M_4), it is not possible for the eavesdropper to detect which packets are received by each destination node. Another interesting observation is that it suffices to make relays M_2, M_4 covert in order to obtain perfect anonymity. This shows that, although making all relays covert ensures perfect secrecy, it may not be necessary.

VII. CONCLUSIONS

One of our key contributions in this work is the theoretical model for anonymity against traffic analysis. To the best of our knowledge, this is the first analytical metric designed to measure the secrecy of *routes* in an eavesdropped wireless

network. Based on the metric, we designed scheduling and relaying strategies to maximize network performance with a guaranteed level of anonymity. Although we consider specific constraints on delay and bandwidth, the ideas of covert relaying and the randomized selection are quite general, and apply to arbitrary multihop wireless networks. The throughput-anonymity tradeoff we obtain reiterates the known paradigm of inverse relationship between communication rate and secrecy in covert channels.

In this work, we used throughput as an indicator of network performance and optimized the selection strategy. However, the framework we establish extends beyond maximizing throughput. In fact, the loss function we define in (18) can be redefined to represent the loss in any convex function of the achievable relay rates. Further, instead of fixing the packet delay and minimizing the loss in sum-rate, we could fix the rates of transmission and analyze the increase in latency at every covert relay. By optimally designing the loss function to reflect the increase in overall network latency, we would be able to derive the relationship between latency and level of anonymity.

APPENDIX

Proof of Theorem 1

To prove the theorem, we adopt the technique used in [20]. Consider the two point processes $\mathcal{Y}_{S_1}, \mathcal{Y}_B$. Let X_j be the j th packet delay, *i.e.* $X_j = Y_B(j) - Y_{S_1}(j)$. Define

$$Z_j \triangleq X_j - X_{j-1} = (Y_{S_1}(j) - Y_B(j-1)) - (Y_{S_1}(j-1) - Y_B(j-1)).$$

We see that Z_j 's are i.i.d. random variables; each Z_j is the difference between two independent exponential random variables with mean $1/C_B$ and $1/C_{S_1}$, respectively. The process $\{X_j\}_{j=1}^\infty$ is a general random walk with step Z_j . Define $X_0 = 0$.

Now for every dummy packet transmitted at t in \mathcal{Y}_B , we insert a virtual packet at t in \mathcal{Y}_{S_1} ; for every packet dropped at time s in \mathcal{Y}_{S_1} , we insert a virtual packet at $s + \Delta$ in \mathcal{Y}_B . Let the new packet delays after the insertion of virtual packets be $\{X'_j\}_{j=0}^\infty$. It can be shown that $\{X'_j\}_{j=0}^\infty$ is also a random walk with step Z_j , but it has two absorbing barriers at 0 and Δ , *i.e.*

$$X'_j = \min(\max(X'_{j-1} + Z_j, 0), \Delta).$$

Since it is almost surely impossible for $X'_{j-1} + Z_j$ to be exactly equal to 0 or Δ , each time $X'_j = 0$ corresponds to a dummy transmission in \mathcal{Y}_B , and $X'_j = \Delta$ corresponds to a dropped packet in \mathcal{Y}_{S_1} . From example 2.16 in [30], we know that the probability of $X'_j = \Delta$ is given by

$$\Pr\{X'_j = \Delta\} = \frac{1 - \frac{T_{S_1}}{T_B}}{\frac{T_B}{T_{S_1}}e^{-\Delta(T_{S_1}-T_B)} - \frac{T_{S_1}}{T_B}} = \Pr\{X'_j = 0\}.$$

Therefore, the fraction of dropped packets in \mathcal{Y}_{S_1} is

$$\epsilon_A = \frac{\Pr\{X'_i = \Delta\}}{(1 - \Pr\{X'_i = 0\})} = \frac{T_B - T_{S_1}}{T_B e^{-\Delta(T_{S_1}-T_B)} - T_{S_1}}.$$

By replacing the transmission rates T_{S_i}, T_B with the maximum values C_{S_i}, C_B , the theorem is proved. In [23], the

authors have shown that the BGM algorithm inserts the least chaff fraction for any pair of point processes. Hence, for any (T_{S_1}, T_B) , it is impossible to obtain a higher information relay rate than (6). This procedure can be extended to multihop by considering multidimensional random walk, but closed form evaluation of the relay rates is cumbersome, even for a few hops. \square

Proof of Theorem 2

2. The outer bound is obtained using the optimality of BGM algorithm. Let node S_i transmit at rates C_{S_i} . Then, the sum information relay rate obtained by using the BGM algorithm on the joint incoming process is given by:

$$\sum_i \lambda_i = (C_{S_1} + C_{S_2})(1 - f_e \left(\sum_i C_{S_i}, C_B \right)). \quad (20)$$

Since BGM inserts the least fraction of dummy packets[23], this is the maximum sum-rate achievable for the given transmission rates. For each individual source S_i , the best rate possible is obtained if the other source is completely ignored. Therefore, by replacing $\sum_j C_{S_j}$ by C_{S_i} in (20), we can obtain the remaining conditions that specify the outer bound. \square

1. Let the zero priority region of Corollary 1 be represented by \mathcal{R}_0 . Every point on the boundary of \mathcal{R}_0 , is obtained by letting one node transmit at the highest rate and varying the transmission rate of the other source node from 0 to the maximum value C_{S_i} . This is a special case of priority mapping; the reduced rate for a node is equivalent to marking a fraction of epochs (in a full rate transmission) to be given equal priority. If we forget about the unmarked epochs, then the rate region is identical to Corollary 1. However the unmarked epochs owing to unused transmissions in the output schedule still have a chance of being relayed and the BGM algorithm can be used between the unmarked epochs of the input and unused epochs of the output. This successive application of BGM amounts to time-sharing between the zero priority and high priority strategies. Since the point on the boundary of \mathcal{R}_0 has a reduced rate of transmission for one node, it is strictly in the interior of priority achievable rate region. Therefore, the bounding convex polygon forms an inner bound to the best achievable rate region. Evaluating the tangents at the maximum sum-rate point of Corollary 1 yield the expressions in Theorem 2. \square

Proof of Theorem 3

Consider the modified point processes as defined in the proof of Theorem 1. X'_i denotes the i^{th} step size of the random walk between two absorbing barriers. The average delay incurred by the BGM algorithm is equal to the expected mean size of the random walk without including the steps that hit either boundaries. Following the exposition in example 2.16 in ([30], Page 67), the cumulative distribution of the step size (or delay Δ_i) in the interval $(0, \Delta)$ is given by

$$\Pr(X_i \leq x) = \frac{1 - \frac{C_{S_1}}{C_B} \exp(\Delta^* + x)(C_{S_1} - C_B)}{1 - \frac{C_{S_1}^2}{C_B^2} \exp(\Delta^*(C_{S_1} - C_B))}. \quad (21)$$

Using the expression above, the average delay $\bar{\Delta}$ for the BGM algorithm with strict delay Δ can be evaluated as:

$$\begin{aligned}\bar{\Delta} &= \mathbb{E}\{X'_i | X'_i \in (0, \Delta^*)\} \\ &= \frac{1 + \exp(\Delta^*(C_{S_1} - C_B)) [\Delta^*(C_{S_1} - C_B) - 1]}{(C_B - C_{S_1}) [1 - \exp(\Delta^*(C_{S_1} - C_B))]}.\end{aligned}$$

If $C_B > C_{S_1}$, then as $\Delta^* \rightarrow \infty$,

$$\begin{aligned}\bar{\Delta} &= \frac{1 + \exp(\Delta^*(C_{S_1} - C_B)) \Delta^*(C_{S_1} - C_B)}{(C_B - C_{S_1}) [1 - \exp(\Delta^*(C_{S_1} - C_B))]} \\ &= \frac{1}{C_B - C_{S_1}}.\end{aligned}$$

This implies that if $\bar{\Delta} > \frac{1}{C_B - C_{S_1}}$, then the BGM algorithm with $\Delta^* = \infty$ would be sufficient, and more importantly, optimal. It is easy to see that for small values of Δ , the average delay $\bar{\Delta} \approx \frac{\Delta^*}{2}$. In other words, when the allowed delay is very small, relaxing the constraint does not provide significant improvement. \square

Proof of Lemma 1

Consider the modified point processes as defined in the proof of Theorem 1. X'_i denotes the i^{th} step size of the random walk between two absorbing barriers. Consider a subsequence \hat{X}_i of X'_i , wherein Z' contains all points in X' that are strictly greater than 0. In other words \hat{X}_i does not represent any dummy packets. Accordingly the erasure variable $E(i) = 1_{0 < \hat{X}_i < \Delta}$ because a packet is relayed whenever the random walk does not hit either barriers. Since the point processes are renewal processes, the resulting random walk is stationary and the distribution for X'_i given by (21). Therefore the erasure $E(i)$ is a stationary and ergodic Markov chain and the capacity of the erasure channel is given by

$$\begin{aligned}\lim_{n \rightarrow \infty} \frac{1}{n} \sum_i E(i) &= 1 - \Pr\{\hat{X}_i = \Delta\} \\ &= 1 - \frac{\Pr\{X'_i = \Delta\}}{(1 - \Pr\{X'_i = 0\})} \\ &= 1 - \frac{1 - \frac{T_{S_1}}{T_B}}{\frac{T_B}{T_{S_1}} e^{-\Delta(T_{S_1} - T_B)} - \frac{T_{S_1}}{T_B}} \\ &= 1 - \epsilon(S_1, B).\end{aligned}$$

\square

Proof of Theorem 4

From (17), we know that $\bar{\lambda}^c(\mathbf{S}, \mathbf{B})$ is an achievable relay rate vector when nodes in \mathbf{B} are covert. It remains to be seen that the condition $H(\mathbf{S}|\hat{\mathbf{S}}) \geq \alpha$ guarantees an anonymity α . For this purpose, it is sufficient to show that

$$H(\mathbf{S}|\mathcal{Y}) \leq H(\mathbf{S}|\hat{\mathbf{S}}).$$

Let $\hat{\mathcal{Y}}$ be the schedules generated assuming $\hat{\mathbf{S}}$ was a session and none of the nodes were covert. The transmission rates of nodes in $\hat{\mathcal{Y}}$ are assumed identical to \mathcal{Y} . For the nodes that are the sources in \mathbf{S} , the schedules are independent in \mathcal{Y} and $\hat{\mathcal{Y}}$. Session $\hat{\mathbf{S}}$ has additional sources due to the broken paths, which also generate independent transmission schedules. The

set of these additional sources is identical to the set of covert relays in \mathbf{S} . Therefore, the schedules are independent in \mathcal{Y} as well. Since the remaining nodes relay all received packets within negligible processing delay, $p(\mathcal{Y}|\mathbf{S}) = p(\hat{\mathcal{Y}}|\mathbf{S})$. Then, using the data processing inequality ($\mathbf{S} - \hat{\mathbf{S}} - \mathcal{Y}$)

$$H(\mathbf{S}|\mathcal{Y}) = H(\mathbf{S}|\hat{\mathcal{Y}}) \leq H(\mathbf{S}|\hat{\mathbf{S}}).$$

\square

Proof of Theorem 5

Consider the optimal solution $q^*(\hat{\mathbf{S}}|\mathbf{S})$ of the distortion rate problem,

$$D = \min_{q(\hat{\mathbf{S}}|\mathbf{S}): I(\mathbf{S}; \hat{\mathbf{S}}) \leq (1-\alpha)H(\mathbf{S})} \mathbb{E}(d(\mathbf{S}, \hat{\mathbf{S}})).$$

From the definition of $d(\mathbf{S}, \hat{\mathbf{S}})$, it is easy to see that if $\nexists \mathbf{B}$ s.t. $\hat{\mathbf{S}} = \mathbf{T}(\mathbf{S}, \mathbf{B})$, then $q^*(\hat{\mathbf{S}}|\mathbf{S}) = 0$. Given $\mathbf{S}, \hat{\mathbf{S}}$, we can show that the set of covert relays \mathbf{B} are uniquely determined, using the following argument:

Suppose $\exists \mathbf{B}_1 \neq \mathbf{B}_2$ such that $\mathbf{T}(\mathbf{S}, \mathbf{B}_1) = \mathbf{T}(\mathbf{S}, \mathbf{B}_2)$. Then, we can write $\mathbf{B}_1 = (\mathbf{B}, \mathbf{B}'_1), \mathbf{B}_2 = (\mathbf{B}, \mathbf{B}'_2)$ where $\mathbf{B}'_1 = (B_{11}, \dots, B_{1m}), \mathbf{B}'_2 = (B_{21}, \dots, B_{2n})$ and $\mathbf{B}'_1 \cap \mathbf{B}'_2 = \emptyset$. We know that

$$\begin{aligned}\hat{\mathbf{S}}(\mathbf{S}, \mathbf{B}_1) &= t(\dots t(\mathbf{T}(\mathbf{S}, \mathbf{B}), B_{11}), \dots), B_{1m}) \\ &= t(\dots t(\mathbf{T}(\mathbf{S}, \mathbf{B}), B_{21}), \dots), B_{2n}) = \hat{\mathbf{S}}(\mathbf{S}, \mathbf{B}_2).\end{aligned}$$

Suppose none of the paths in $\mathbf{T}(\mathbf{S}, \mathbf{B})$ contain $\mathbf{B}'_1 \cup \mathbf{B}'_2$, then it does not matter if those relays are covert or not, in which case the subset of covert relays would be \mathbf{B} .

If $\exists P \in \mathbf{T}(\mathbf{S}, \mathbf{B})$ that contains B_{11} , then $\mathbf{T}(\mathbf{S}, \mathbf{B}_1)$ would contain a path that ends in B_{11} , whereas $\mathbf{T}(\mathbf{S}, \mathbf{B}_2)$ cannot contain such a path. Therefore, we have a contradiction.

The above argument shows that we can equivalently write $q^*(\hat{\mathbf{S}}|\mathbf{S}) = q^*(\mathbf{B}|\mathbf{S})$. Therefore, q^* specifies a valid selection strategy. Since $H(\mathbf{S})$ is fixed apriori, $I(\mathbf{S}; \hat{\mathbf{S}}) \leq (1-\alpha)H(\mathbf{S})$ ensures that an anonymity α is guaranteed. Further, for every \mathbf{B} , the function d evaluates the difference in achievable rate vectors $\bar{\lambda}^v(\mathbf{S})$ and $\bar{\lambda}^c(\mathbf{S}, \mathbf{B})$. Taking expectation over $q^*(\mathbf{B}|\mathbf{S})$, it is easy to see that the distortion D is achievable with α -anonymity. \square

Switching Network Example

When all relays are visible, the eavesdropper would not know the final node of any route. This implies that given an observation, 4 possible source-destination pairings would be equally likely. This implies that his uncertainty $H(\mathbf{S}|\mathcal{Y}) = \log(4)$. Since the priors are equally likely $H(\mathbf{S}) = \log(24)$. Therefore, when all relays are visible, $\alpha = \frac{\log(4)}{\log(24)} = .436$.

When M_1, M_3 are covert, the number of possible pairings given an observation would depend on the session. For example, if $\{(S_1, M_1, M_2, D_1), (S_2, M_1, M_2, D_2), (S_3, M_3, M_4, D_3), (S_4, M_3, M_4, D_4)\}$ is the session, then the eavesdropper would be able to identify that all transmissions from M_1 are relayed by M_2 , and his uncertainty would

be $\log(4)$. This is identical to 7 other pairings (whenever S_1, S_2 use the same set of relays). Suppose $\{(S_1, M_1, M_2, D_1), (S_2, M_1, M_4, D_3), (S_3, M_2, M_3, D_2), (S_4, M_2, M_4, D_4)\}$ was the session, then it would be indistinguishable from the 15 remaining sessions (whenever S_1, S_2 do not use the same set of relays), and his uncertainty would increase to $\log(16)$. Therefore, since all sessions are equally probable,

$$\frac{H(\mathbf{S}|\mathcal{Y})}{H(\mathbf{S})} = \frac{(1/3)\log(4) + (2/3)\log(16)}{\log(24)} = 0.659.$$

REFERENCES

- [1] N. West, *The SIGINT Secrets: The Signal Intelligence War: 1900 to Today*. New York: William Morrow, 1988.
- [2] V. L. Vodyack and S. T. Kent, "Security mechanisms in high-level network protocols," *ACM Computing Surveys*, vol. 15, pp. 135–171, 1983.
- [3] J.-F. Raymond, "Traffic analysis: Protocols, attacks, design issues and open problems," in *Designing Privacy Enhancing Technologies: Proceedings of International Workshop on Design Issues in Anonymity and Unobservability* (H. Federrath, ed.), vol. 2009 of LNCS, pp. 10–29, Springer-Verlag, 2001.
- [4] Q. Sun, D. R. Simon, Y. Wang, W. Russell, V. N. Padmanabhan, and L. Qiu, "Statistical identification of encrypted web browsing traffic," in *Proceedings of the 2002 IEEE Symposium on Security and Privacy*, (Berkeley, California), p. 19, May 2002.
- [5] N. Mathewson and R. Dingledine, "Practical traffic analysis: Extending and resisting statistical disclosure," in *Privacy Enhancing Technologies: 4th International Workshop*, May 2004.
- [6] E. W. Felten and M. A. Schneider, "Timing attacks on web privacy," in *ACM Conference on Computer and Communications Security*, pp. 25–32, 2000.
- [7] D. X. Song, D. Wagner, and X. Tian, "Timing Analysis of Keystrokes and Timing Attacks on SSH," in *Proc. 10th USENIX Security Symposium*, (Washington, DC), August 2001.
- [8] C. E. Shannon, "Communication theory of secrecy systems," *Bell System Technical Journal*, 1949.
- [9] D. Chaum, "Untraceable electronic mail, return addresses and digital pseudonyms," *Communications of the ACM*, vol. 24, pp. 84–88, February 1981.
- [10] D. Kesdogan, J. Egner, and R. Buschkes, "Stop-and-go MIXes providing probabilistic security in an open system," in *Second International Workshop on Information Hiding (IH'98), Lecture Notes in Computer Science*, vol. 1525, (Portland, Oregon), pp. 83–98, April 1998.
- [11] A. Pfitzmann, B. Pfitzmann, and M. Waidner, "ISDN-MIXes: Untraceable communication with very small bandwidth overhead," in *Proceedings of the GI/ITG Conference: Communication in Distributed Systems, Informatik-Fachberichte*, vol. 267, (Mannheim, Germany), pp. 451–463, February 1991.
- [12] C. Gulcu and G. Tsudik, "Mixing e-mail with babel," in *Proceedings of the Symposium on Network and Distributed System Security*, pp. 2–19, February 1996.
- [13] M. K. Reiter and A. D. Rubin, "Crowds: anonymity for Web transactions," *ACM Transactions on Information and System Security*, vol. 1, no. 1, pp. 66–92, 1998.
- [14] G. Danezis, R. Dingledine, and N. Mathewson, "Mixminion: design of a type iii anonymous remailer protocol," in *Proceedings of 2003 Symposium on Security and Privacy*, pp. 2–15, May 2003.
- [15] Y. Zhu, X. Fu, B. Graham, R. Bettati, and W. Zhao, "On flow correlation attacks and countermeasures in mix networks," in *Proceedings of Privacy Enhancing Technologies workshop*, May 26–28 2004.
- [16] B. Radosavljevic and B. Hajek, "Hiding traffic flow in communication networks," in *Military Communications Conference*, 1992.
- [17] A. Wyner, "The wiretap channel," *Bell Syst. Tech. J.*, vol. 54, pp. 1355–1387, 1975.
- [18] I. Csiszár and J. Körner, "Broadcast channels with confidential messages," *IEEE Trans. on Information Theory*, vol. 24, pp. 339–348, May 1978.
- [19] S. Axelsson, "Intrusion detection systems: A taxonomy and survey," tech. rep., Chalmers University of Technology, Sweden, March 2000.
- [20] T. He and L. Tong, "Detecting Information Flows: Fundamental Limits and Optimal Algorithms," submitted to *IEEE Trans. on Information Theory*, 2007.
- [21] T. Cover and J. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 1991.
- [22] A. Serjantov and G. Danezis, "Towards an information theoretic metric for anonymity," in *Proceedings of Privacy Enhancing Technologies Workshop (PET 2002)* (R. Dingledine and P. Syverson, eds.), Springer-Verlag, LNCS 2482, April 2002.
- [23] A. Blum, D. Song, and S. Venkataraman, "Detection of Interactive Stepping Stones: Algorithms and Confidence Bounds," in *Conference of Recent Advance in Intrusion Detection (RAID)*, (Sophia Antipolis, French Riviera, France), September 2004.
- [24] S. Boucheron and M. R. Salamatian, "About Priority Encoding Transmission," *IEEE Transactions on Information Theory*, vol. 46, pp. 699–705, March 2000.
- [25] N. Shacham and P. McKenney, "Packet Recovery in High-Speed Networks using Coding and Buffer Management," in *Proc. IEEE INFOCOM*, pp. 124–131, 1990.
- [26] L. Rizzo, "Effective Erasure Codes for Reliable Computer Communication Protocols," in *Proc. ACM SIGCOMM Computer Communication Review*, vol. 27, pp. 24–36, 1997.
- [27] T. He, P. Venkatasubramanian, and L. Tong, "Packet scheduling against stepping-stone attacks with chaff," in *Proc. IEEE Military Communications Conference*, (Washington, DC), October 2006.
- [28] R. Blahut, "Computation of Channel Capacity and Rate-Distortion Functions," *IEEE Trans. Infor. Theory*, vol. IT-18, July 1972.
- [29] D. Neuhoff and L. Gilbert, "Causal Source Codes," *IEEE Trans. on Information Theory*, vol. 28, pp. 701–713, Sep. 1982.
- [30] D. Cox and H. Miller, *The Theory of Stochastic Processes*. New York: John Wiley & Sons Inc., 1965.